基于支撑函数的概念格属性约简*)

葛方斌[1] 杨 林[2] 刘 伟[1] 王建新[2]

(解放军理工大学指挥自动化学院 南京 210007)[1] (中国电子系统工程研究所 北京 100039)[2]

摘 要 依据概念格中一个概念的内涵属性相对于另一个概念的内涵的不同关系,提出了区分属性概念,并研究了其性质;利用区分属性构造出概念格支撑函数。提出并证明了以支撑函数为基础的形式背景属性约简定理,改进了文献[7]中基于辨识函数的属性约简方法。

关键词 形式背景,概念格,区分属性,支撑函数,属性约简

Improvement of Attributes Reduction of Concept Lattices

GE Fang-bin[1] YANG Lin[2] LIU Wei[1] WANG Jian-xin[2]

(College of Command Automation, PLA University of Science and Technology, Nanjing 210007, China)[1]

(Institute of China Electronic System Engineering, Beijing 100039, China)[2]

Abstract There are two kinds of relations between the connotation attributes of concepts and the connotation of another concept in concept lattices. Based on them, the concept of distinguishable attributes is presented, and its properties are studied. At the same time, the supporting function is constructed with distinguishable attributes, and the theorem, which is about attributes reduction of formal contexts, is presented and proved based on supporting functions. The method of attributes reduction in the literature [7] is improved.

Keywords Formal context, Concept lattice, Distinguishable attribute, Supporting function, Attributes reduction

概念格[1]由 Will. R 于 1982 年提出。它是基于形式背景中对象与属性间的二元关系构造出的一种层次性概念结构,精确地刻画了概念间的泛化与特化关系。作为数据处理和知识发现的形式化工具,概念格模型已在数据挖掘、软件工程、知识工程等众多领域[3-6]得到成功应用。概念格中概念属性约简是概念性知识简化的重要手段,文[7]提出一种以辨识函数为基础的属性约简方法。该方法需要针对每两个概念生成一个布尔表达式,构造出的辨识函数一般很复杂。本文在充分研究区分属性特性的基础上,提出了支撑函数概念,将支撑函数用于概念格属性约简。支撑函数的构造只需要相邻概念间的区分属性,因而其形式结构远比辨识函数来得简单,所以基于支撑函数的属性约简明显优于基于辨识函数的属性约简。

1 形式背景与概念格

概念格理论也称形式概念分析[1,2],概念格以形式背景为基础,形式背景的定义为:

定义 1[2] 三元组 $K=(U,D,R)$称为一个形式背景。其中,U 是非空有限的对象集合,D 是有限的属性集合,R 是 U 和 D 间的二元关系,即 $R\subseteq U\times D$。若$(x,a)\in R$,则称对象 x 具有属性 a,记为 xRa。

在形式背景(U,D,R)中,对 $\forall x\in U$, $\forall a\in D$,若用 1 表示$(x,a)\in R$,0 表示$(x,a)\notin R$,则形式背景可表示为只有 0 和 1 的表格。

对形式背景(U,D,R),定义 U,D 幂集上的映射 $f:2^U\to 2^D$,$g:2^D\to 2^U$ 如下:

$$f(X)=\begin{cases}\{a\mid a\in D\wedge\forall x\in X(xRa)\}\\ D\end{cases}$$

$X\in 2^U, X\neq\varphi$

$X=\varphi$

$$g(A)=\begin{cases}\{x\mid x\in U\wedge\forall a\in A(xRa)\}\\ U\end{cases}$$

$A\in 2^D, A\neq\varphi$

$A=\varphi$

任意 $x\in U$,若 $f(\{x\})\neq\varphi$,则称(U,D,R)是 U 上正则的;任意 $a\in D$,若 $g(\{a\})\neq\varphi$,则称(U,D,R)是 D 上正则的;若(U,D,R)既是 U 上正则的又是 D 上正则的,则称(U,D,R)是正规的。

定义 2[2] 设(U,D,R)为形式背景,如果一个二元组(X,A)满足 $X\subseteq U, A\subseteq D$, $f(X)=A, g(A)=X$,则称(X,A)是一个形式概念,简称概念。其中,X 称为概念的外延,A 称为概念的内涵,A 中属性称为概念(X,A)的特征属性。

显然,任意 $X\subseteq U$,$(X,f(X))$是概念,当且仅当 $g(f(X))=X$;任意 $A\subseteq D$,$(g(A),A)$是概念,当且仅当 $f(g(A))=A$。

设 $L(U,D,R)$为形式背景(U,D,R)中所有概念构成的集合,定义 $L(U,D,R)$上的序关系:

$(X_1,A_1)\leqslant(X_2,A_2)\Leftrightarrow X_1\subseteq X_2(\Leftrightarrow A_1\supseteq A_2)$

偏序集 $L(U,D,R)$按如下定义的上、下确界构成一个格,称其为概念格。$L(U,D,R)$还是一个完备格[2]。

$(X_1,A_1)\wedge(X_2,A_2)=(X_1\cap X_2, f(g(A_1\cup A_2)))$

*)项目资助:国家"863"高技术研究发展计划基金项目(2005AA147050)。**葛方斌** 博士研究生,主要研究方向为计算机网络、信息安全;**杨 林** 博导,研究员,主要研究方向为计算机网络系统、信息安全;**刘 伟** 博士,研究生,主要研究方向为信息安全;**王建新** 博导,研究员,主要研究方向为指挥自动化理论、信息安全。

$(X_1,A_1)\vee(X_2,A_2)=(g(f(X_1\cup X_2)),A_1\cap A_2)$

概念格中的概念具有下列基本性质[2]($\forall X_1,X_2,X\subseteq U,\forall A_1,A_2,A\subseteq D$)

(1)$X_1\subseteq X_2\Rightarrow f(X_2)\subseteq f(X_1),A_1\subseteq A_2\Rightarrow g(A_2)\subseteq g(A_1)$

(2)$X\subseteq g(f(X)),A\subseteq f(g(A))$

(3)$f(X)=f(g(f(X))),g(A)=g(f(g(A)))$

(4)$X\subseteq g(A)\Leftrightarrow A\subseteq f(X)$

(5)$f(X_1\cup X_2)=f(X_1)\cap f(X_2),g(A_1\cup A_2)=g(A_1)\cap f(A_2)$

(6)$f(X_1\cap X_2)\supseteq f(X_1)\cup f(X_2),g(A_1\cap A_2)\supseteq g(A_1)\cup f(A_2)$

(7)$(g(f(X)),f(X)),(g(A),f(g(A)))$都是概念。

2 区分属性及其性质

形式背景中的任意两个外延不同的概念,它们的内涵必然也是不同的。也就是说必有其中一个概念存在某个特征属性不是另一个概念的特征属性,该属性形成两个概念的一个区分。定义3以另一种方式对区分属性进行了界定。

定义3 设(U,D,R)是一形式背景,$(X,A),(Y,B)\in L(U,D,R),a\in A$,若$Y\not\subseteq g(\{a\})$,则称$a$是$(X,A)$关于$(Y,B)$的区分属性,否则,称$a$是$(X,A)$关于$(Y,B)$的非区分属性。

根据概念的性质,$(g(\{a\}),f(g(\{a\})))$也是(U,D,R)中的概念,称其为$\{a\}$生成的概念。a是(X,A)关于(Y,B)的区分属性说明,$\{a\}$的生成概念$(g(\{a\}),f(g(\{a\})))$不是(Y,B)的泛化,即$(Y,B)\not\leqslant(g(\{a\}),f(g(\{a\})))$。相应地,$(X,A)$关于$(Y,B)$的非区分属性就是$(X,A)$的特征属性中生成概念是$(Y,B)$泛化的属性。

定理1 设(U,D,R)是一形式背景,$(X,A),(Y,B)\in L(U,D,R),a\in A$,则$a$是$(X,A)$关于$(Y,B)$的区分属性,当且仅当$a\notin B$。

证明:充分性。反设a是(X,A)关于(Y,B)的非区分属性,则$Y\subseteq g(\{a\})$,从而$f(g(\{a\}))\subseteq f(Y)$;加之$a\in\{a\}\subseteq f(g(\{a\}))$,$f(Y)=B$,于是,$a\in B$,这与前提矛盾。

必要性。反设$a\in B$,则$Y=g(B)\subseteq g(\{a\})$;加之$a\in A$,从而a是(X,A)关于(Y,B)的非区分属性,这与前提矛盾。

由上述区分属性的性质可知,概念(X,A)关于概念(Y,B)的区分属性集为$A-B$。

下面的推论是显然的。

推论 设(U,D,R)是一形式背景,$(X,A),(Y,B),(Z,C)\in L(U,D,R)$,$(Y,B)\leqslant(Z,C)$,若$a$是$(X,A)$关于$(Y,B)$的区分属性,则$a$也是$(X,A)$关于$(Z,C)$的区分属性。

定理2 设(U,D,R)是一形式背景,$(X,A),(Y,B)\in L(U,D,R)$,若$C\subseteq D$包含(X,A)关于(Y,B)的区分属性,则$Y\not\subseteq g(C)$。

证明:假设C包含(X,A)关于(Y,B)的区分属性a。反设$Y\subseteq g(C)$。由$a\in C$有,$g(C)\subseteq g(\{a\})$,从而$Y\subseteq g(\{a\})$,这与a是(X,A)关于(Y,B)的区分属性矛盾。

区分属性的定义说明,概念(X,A)关于概念(Y,B)的区分属性的生成概念不可能是(Y,B)的泛化,定理2进一步指出,任何包含(X,A)关于(Y,B)的区分属性的属性集,其生成概念同样不可能是(Y,B)的泛化。

3 概念格属性约简

形式背景的属性约简以概念格同构为基础,目的是在保持形式背景概念构造能力不变的前提下找出最简属性子集,是概念性知识简化的重要手段。

定义4 设$(X_1,A_1),(X_2,A_2)$分别是$L(U,D_1,R_1),L(U,D_2,R_2)$中的概念,若$X_1=X_2$,则称$(X_1,A_1)$与$(X_2,A_2)$相似,记为$(X_1,A_1)\triangleq(X_2,A_2)$。若对于任意$(X,A)\in L(U,D_2,R_2)$,都存在$(X',A')\in L(U,D_1,R_1)$,使$(X',A')\triangleq(X,A)$,则称$L(U,D_1,R_1)$细于$L(U,D_2,R_2)$,记为

$L(U,D_1,R_1)\leqslant L(U,D_2,R_2)$

若$L(U,D_1,R_1)\leqslant L(U,D_2,R_2)$,且$L(U,D_2,R_2)\leqslant L(U,D_1,R_1)$,则称$L(U,D_1,R_1)$与$L(U,D_2,R_2)$同构[7],记为

$L(U,D_1,R_1)\cong L(U,D_2,R_2)$

在形式背景(U,D,R)下,对任意$A\subseteq D$,定义$R_A=R\cap(U\times A)$,则(U,A,R_A)也是一个形式背景。$L(U,A,R_A)$和$L(U,D,R)$有下面的关系。

定理3[7] 设(U,D,R)是一形式背景,$A\subseteq D$,则$L(U,D,R)\leqslant L(U,A,R_A)$。

$L(U,D,R)\leqslant L(U,A,R_A)$显示,任何$D$的子集在针对$U$的概念构造能力方面是不强于$D$的。$D$的概念构造能力不减的最简子集以尽可能少的属性完整地刻画了原有形式背景中的知识,这样的属性子集就是属性约简。其严格定义为:

定义5 设(U,D,R)是一形式背景,$A\subseteq D$,若$L(U,A,R_A)\cong L(U,D,R)$,且$\forall A'\subset A,L(U,A',R_{A'})\ncong L(U,D,R)$,则称$A$是$(U,D,R)$的一个约简。

对形式背景(U,D,R),由于$U\subseteq g(f(U))$,且$g(f(U))\subseteq U$,可见,$(U,f(U))$一定是$L(U,D,R)$中的概念,称其为平凡概念。任何形式背景中都存在平凡概念,只含有平凡概念的概念格称为平凡概念格。

定理4 设(U,D,R)是一形式背景,则(U,D,R)的约简存在,当且仅当$L(U,D,R)$是非平凡概念格。

证明:充分性。由于$L(U,\varphi,R_\varphi)$是平凡概念格,因此$L(U,\varphi,R_\varphi)$与$L(U,D,R)$不同构。考虑D的所有一元子集,若存在$a\in D$,使$L(U,\{a\},R_{\{a\}})\cong L(U,D,R)$,则$\{a\}$显然是$(U,D,R)$关于$L(U,D,R)$的一个约简。否则,考虑$D$的所有二元子集,若存在二元子集$A\subseteq D$,使$L(U,A,R_A)\cong L(U,D,R)$,则$A$是$(U,D,R)$关于$L(U,D,R)$的一个约简。否则考虑$D$的所有三元子集,如此进行下去。由于$D$是有限集合,且$L(U,D,R)$与自身同构,故在有限步内必可找到$(U,D,R)$的一个约简,故$(U,D,R)$的约简存在。

必要性。若(U,D,R)的约简存在,不妨设A是(U,D,R)的一个约简。由约简的定义知,存在$B\subseteq A$,使,$L(U,B,R_B)$与$L(U,D,R)$不同构,又由于$L(U,D,R)\leqslant L(U,B,R_B)$,从而$L(U,D,R)$中一定包含非平凡概念,故$L(U,D,R)$是非平凡概念格。

形式背景(U,D,R)中,若$(X,A),(Y,B)\in L(U,D,R),X\neq Y$,且$(X,A)\leqslant(Y,B)$,则称$(Y,B)$是$(X,A)$的后继。进一步,对$\forall(Z,C)\in L(U,D,R)$,若$(X,A)\leqslant(Z,C)\leqslant(Y,B)$,都有$Z=X$或$Z=Y$,则称$(Y,B)$是$(X,A)$的直接后继。

对非平凡概念(X,A),令$S_{(X,A)}=\{(Y,B)\mid(Y,B)$是$(X,A)$的直接后继$\}$。对每个$a\in D$,指定一个布尔变量,仍用"$a$"不加区分地表示。利用$(X,A)$关于其每个直接后继的区分属性构造一个与$(X,A)$对应的布尔表达式$\Delta_{(X,A)}$。

$$\Delta_{(X,A)}=\bigwedge_{(Y,B)\in S_{(X,A)}}\left(\bigvee_{a\in A-B}a\right)$$

依据概念格中所有非平凡概念(X,A)对应的$\Delta_{(X,A)}$可定

义概念格的支撑函数。

定义 6 在形式背景(U,D,R)中，若

$$\Delta=\bigwedge_{(X,A)\in L(U,D,R)\atop X\neq U}\Delta_{(X,A)}$$

则称Δ是$L(U,D,R)$的支撑函数。

$L(U,D,R)$的支撑函数是一布尔合取范式，它可化为极小析取范式。此极小析取范式直接确定了形式背景的所有约简。

定理 5 A是形式背景(U,D,R)的一个约简，当且仅当A是Δ的极小析取范式的一个析取支。

证明：充分性。首先证明，$L(U,A,R_A)\cong L(U,D,R)$。根据定理 3，只需证明，$\forall(X,B)\in L(U,D,R)$，$L(U,A,R_A)$中存在(X,B)的相似概念。若$X=U$，结论显然成立；若$X\neq U$，则对$L(U,D,R)$中任意(X,B)的后继(Y,C)，当(Y,C)是(X,B)的直接后继时，$B-C$构成Δ的一合取支，由于A是Δ的极小析取范式的一个析取支，从而$A\cap(B-C)\neq\varphi$，令$A\cap B=E$，则$E\cap(B-C)=(A\cap B)\cap(B-C)=A\cap(B\cap(B-C))=A\cap(B-C)\neq\varphi$，由定理 1 知，$B-C$中的属性都是$(X,B)$关于$(Y,C)$的区分属性，$E\cap(B-C)\neq\varphi$说明，$E$包含$(X,B)$关于$(Y,C)$的区分属性。根据定理 2 有，$g(E)\neq Y$；当$(Y,C)$不是$(X,B)$的直接后继时，$(Y,C)$显然存在一个直接后继$(Z,F)$，使$(X,B)\leqslant(Z,F)\leqslant(Y,C)$，由前面的证明知，$E$包含$(X,B)$关于$(Z,F)$的区分属性，结合定理 1 的推论知，$E$包含$(X,B)$关于$(Y,C)$的区分属性，从而仍有$g(E)\neq Y$，这说明$(U,A,R_A)$中的概念$(g(E),f(g(E)))$与$(U,D,R)$中$(X,B)$的任何后继概念都不相似，但由定理 3 知，$(U,D,R)$中存在$(g(E),f(g(E)))$的相似概念，另外，由$E=A\cap B\subseteq B$又有，$X=g(B)\subseteq g(E)$，这说明$(U,D,R)$中与$(g(E),f(g(E)))$相似的概念都不小于$(X,B)$，这样必有$g(E)=X$，故$(g(E),f(g(E)))$就是$(U,A,R_A)$中$(X,B)$的相似概念。其次证明，$\forall A'\subset A$，$L(U,A',R_{A'})\ncong L(U,D,R)$。因为$A$是$\Delta$的极小析取范式的一个析取支，所以存在$\Delta$的某个合取支，$A'$不含有该合取支的任何属性。否则，$\Delta$的极小析取范式必有某析取支包含于$A'$，这与$A$是$\Delta$的极小析取范式的一个析取支矛盾。不妨设该合取支是由概念(X,B)与其直接后继(Y,C)确定的$\bigvee_{a\in B-C}a$。于是，$A'\cap(B-C)=\varphi$。令$A'\cap B=E$，则$E-A'\cap C=\varphi$，从而$E\subseteq A'\cap C\subseteq C$，故$g(C)\subseteq g(E)$。这样，在$(U,A',R_{A'})$中，一方面，$f(X)=B\cap A'=E$；另一方面，由$Y=g(C)\subseteq g(E)$，且$X\subset Y$可推知$g(E)\neq X$，由此可见，$(X,f(X))$不是$(U,A',R_{A'})$中的概念，可见$(U,A',R_{A'})$中不存在$(X,B)$的相似概念，故$L(U,A',R_{A'})\ncong L(U,D,R)$。综上可知，$A$是形式背景$(U,D,R)$的一个约简。

必要性。因为A是形式背景(U,D,R)的一个约简，所以A必包含Δ的每个合取支的至少一个属性。否则，由已证充分性可知，若A不含Δ的某个合取支的任何元素，则必有$L(U,A,R_A)\ncong L(U,D,R)$，这与$A$是$(U,D,R)$的约简矛盾。于是，$\Delta$的极小析取范式必存在某析取支$C$，$C\subseteq A$。实际上，$C=A$。若$C\subset A$，因为$A$是$(U,D,R)$的一个约简，所以$L(U,C,R_C)\ncong L(U,D,R)$，从而$C$不是$(U,D,R)$的约简，这与已证充分性矛盾。$A=C$说明$A$就是$\Delta$的极小析取范式的一个析取支。

定理 5 说明，形式背景的所有属性约简可通过相应概念格支撑函数的布尔运算完全确定，这种确定方法的优点是简单直观，而且易于在机器上实现，缺点是计算时间复杂度较高，无法处理大量对象属性构成的复杂概念格属性约简。对于一个不太复杂的概念格而言却不失为一个不错的方法。

4 约简举例

下面借用文献[7]中的例子说明定理的应用。

例：针对表 1 所给形式背景(U,D,R)，求其属性约简。

表 1

	a	b	c	d	e
1	1	1	0	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	1	0	0

(U,D,R)上的概念格为：

$L(U,D,R)=\{(\varphi,D),(1,abde),(24,abc),(13,d),(124,ab),(U,\varphi)\}$

$L(U,D,R)$的支撑函数为：

$$\begin{aligned}L(U,D,R)&=\Delta_{(\varphi,D)}\wedge\Delta_{(1,abde)}\wedge\Delta_{(24,abc)}\wedge\Delta_{(13,d)}\wedge\Delta_{(124,ab)}\\&=[c\wedge(d\vee e)]\wedge[(a\vee b\vee e)\wedge c]\wedge[c]\wedge[d]\wedge[a\vee b]=cd(a\vee b)=acd\vee bcd\end{aligned}$$

可见，(U,D,R)有两个约简，分别是$\{a,c,d\}$，$\{b,c,d\}$。

上述结果与文献[7]中用辨识函数方法求得的结果是一致的，但和文献[7]中的辨识函数相比，例子中的支撑函数形式简洁了许多，这大大简化了约简的计算。正如前面提到的，辨识函数的构造使用了所有概念对的区分属性，而支撑函数的构造只利用大小相邻概念对的区分属性。在形式结构上，支撑函数只是辨识函数的一小部分，在概念格属性约简中用支撑函数取代辨识函数其优点是非常明显的。

结束语 本文依据概念格中概念的特征属性相对于其它概念内涵的不同关联提出了区分属性概念，并在仔细分析了区分属性性质的基础上定义了概念格支撑函数概念。支撑函数由大小相邻概念对的区分属性构成，这些区分属性体现了与其相联系概念间的差别。它们的结合使概念格的层次结构得到保持。这正是支撑函数能够用于属性约简的根本基础。和辨识函数相比，支撑函数在约简方面的优势是明显的，但复杂概念格属性约简仍是该方法面临的最大挑战，是一个值得深入研究的问题。

参考文献

[1] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts//Rival I, ed. Ordered Sets. Dordrecht-Boston: Reidel, 1982:445-470

[2] Gander B, Wille R. Formal Concept Analysis, Mathematical Foundations. Berlin: Springer, 1999

[3] Oosthuizen G D. The Application of Concept Lattice to Machine Learning. Technical Report. University of Pretoria, South Africa, 1996

[4] Godin R, Mineau G W, Missaoui R. incremental structuring of knowledge bases // Proc. International Symposium on Knowledge Retrieval, Use, and Storage for Efficiency (KRUSE'95). Santa Cruz, 1995:179-193

[5] Oosthuizen G D. Rough sets and concept lattices. In: Ziarko W P, ed. Rough Sets, and Fuzzy Sets and Knowledge Discovery (RSKD'93). London: Springer-Verlag, 1994:24-31

[6] Siff M, Reps T. Identifying modules via concept analysis//Harrold M J, Visaggio G, eds. International Conference on Software Maintenance. Bari, Italy. Washington, DC: IEEE Computer Society, 1997: 170-179

[7] 张文修，魏玲，祁建军. 概念格属性约简理论与方法. 中国科学 E 辑，信息科学，2005，36(6)：628-639